NPL Report CISE 42/99

A DISCUSSION OF APPROACHES FOR DETERMINING A REFERENCE VALUE IN THE ANALYSIS OF KEY-COMPARISON DATA

BY
M G COX

October 1999

A discussion of approaches for determining a reference value in the analysis of key-comparison data*

M G Cox

Centre for Information Systems Engineering

October 1999

ABSTRACT

When analysing a sample statistically, the quality of the resulting parameters (measures of location, dispersion, etc.) depends on the reliability or credibility of the data and on the analysis undertaken. The data provided by interlaboratory comparisons, and especially key comparisons, is of sufficient importance that approaches are required which are as defensible and robust as possible. Central to the area of key comparisons is the degree of equivalence of measurement standards, which is the degree to which the standards are consistent with reference values determined from the key comparisons and hence are consistent with one another. This paper indicates some initial thinking at the National Physical Laboratory on methods for determining key-comparison reference values. It is based on the concept of making appropriate assertions relating to the measurement standards provided in a key comparison and devising and utilizing statistical techniques that are consistent with them. The concept is illustrated by applying it to data from a spectral-irradiance key comparison. The report does not attempt to advocate a specific solution, but to convey what is believed to be an appropriate attitude to data analysis in this area.

*To appear in the proceedings of Advanced Mathematical and Computational Tools for Metrology, a conference held in Oxford, UK, 12-15 April 1999, edited by P. Ciarlini, A. B. Forbes, F. Pavese and D. Richter, World Scientific, Singapore.

ISSN 1361-407X

Authorised by Dr Dave Rayner,
Head of the Centre for Information Systems Engineering

National Physical Laboratory, Queens Road, Teddington, Middlesex, TW11 0LW

Contents

1. Introduction

In the traditional analysis of a sample, statistical estimators such as the mean, standard deviation, etc., are determined and used to infer information about the parent population from which the sample is assumed to have been drawn. In particular, a confidence interval associated with an estimator is often required. In determining a confidence interval, an assumption, such as normality, is made about the error distribution of the parent population.

Some samples may be such that the use of a conventional analysis will provide misleading results, perhaps because the data contains "outliers", because an assumption concerning the distribution is invalid, or for some other reason. An alternative approach is therefore today more frequently entertained which uses "robust" or non-parametric techniques in an attempt to overcome such difficulties.

For data obtained from an interlaboratory comparison it is indeed possible to use "traditional" statistics as the basis of quantifying the degree of equivalence of laboratories. It can be expected that there will be particular comparisons for which this approach is satisfactory; for others it will be more valid to use robust or non-parametric approaches.

The technical basis of an agreement (BIPM, 1999; Thomas, 1999) for the mutual recognition of national standards and of calibration and measurement certificates issued by national metrology institutes, drawn up by the Comité International des Poids et Mesures (CIPM), is the results of *key comparisons* of national measurement standards[1]. A key comparison (BIPM, 1999) is one of the set of comparisons carried out by Consultative Committees of the CIPM to test the principal techniques and methods in the field.

Central to the considerations is the concept of the degree of equivalence of measurement standards:

> "*CIPM key comparisons lead to reference values, known as key comparison reference values … the term degree of equivalence of measurement standards is taken to mean the degree to which a standard is consistent with the key comparison reference value. The degree of equivalence of each national measurement standard is expressed quantitatively by two terms: its deviation from the key comparison reference value and the uncertainty of this deviation (at a 95% level of confidence). The degree of equivalence between pairs of national measurement standards is expressed by the difference of their deviations from the reference value and the uncertainty of this difference (at a 95% level of confidence).*" (BIPM, 1999)

Thus, the concept of a key comparison reference value (KCRV), and its accompanying uncertainty, is essential in determining the degree of equivalence of each national measurement standard. Therefore, at NPL, initial attention has concentrated on approaches to determining KCRVs which are as objective and defensible as pragmatically possible.[2]

---

[1] There are currently about one hundred key comparisons.

[2] Other aspects under consideration include experimental-design issues as part of establishing protocols for interlaboratory comparisons which attempt to lead to results in as reliable a form as reasonably possible. They also include the considerations listed in Section 8 of this report.

The main thesis of the work on which this report is based is that any approach for determining a KCRV should be based on agreed and explicit assumptions. In particular, it is regarded as essential that an approach be driven by the metrology area rather than by abstract statistical considerations such as the assumption that normal (Gaussian) statistics can be applied to the values provided by participating laboratories. The approach may well be different for different disciplines and indeed for different comparisons within a discipline. Nevertheless, there are advantages in seeking approaches having generic properties, and considering their tailoring where necessary to specific instances.

For interlaboratory-comparison data the concept of a *homogeneous population* is a dubious one. In a sample, say, of the lifetimes of light bulbs of a particular type manufactured under controlled conditions, it is reasonable to analyse a random sample from a batch and infer properties of the batch as a whole. For interlaboratory-comparison data it is less reasonable to regard the measurement results as a random sample from a population, because of the possibly disparate nature of the measurement methods used to obtain it. Even though careful protocols may be promulgated by the pilot laboratory in supervising the comparison, there is no guarantee that all measurements are carried out totally in accordance with them. Moreover, in a sense, the sample itself *is* the population (or a large part of it), since there may be no other laboratory (or perhaps only a small number of laboratories) which would have the capability to provide the particular measurement standard.

It is therefore desirable to move away from some of the traditional ways of thinking and try to devise approaches which have sensible properties for comparison data, which for the above reasons often has heterogeneous origin. To pre-empt some of the following discussion, one assumption might be that the nature of the approaches used by the laboratories participating in a comparison implies that the errors in their measurement standards can be regarded as independent. In another situation, a group of the laboratories involved might have a degree of commonality, perhaps in terms of their use of a common reference standard, and therefore the measurement errors of this group are correlated. Within this, it may or it may not be possible to quantify the extent of this correlation. It may be possible to assume that the measurement standard of a particular laboratory can be expected to have an error which is equally likely to be positive or negative, i.e., that its measurement standard has the same chance of lying on the low or the high side. Such are some of the considerations that should be taken into account in the analysis of comparison data. Even from meagre information, it is often possible to deduce in a logical way a suitable statistical treatment.

A politically-sensitive issue concerns the credibility that can be attached to the measurement uncertainty assigned by each participating laboratory. For any one particular comparison, most and perhaps all of the participating laboratories can be expected to provide realistic measurement uncertainties. However, in some circumstances there may be reason to believe that not all the participating laboratories provide realistic uncertainties. Evidence for such belief may be indicated by the fact that the set of 95% confidence intervals associated with the measurement standards do not “overlap sufficiently”. The reasons for this lack of agreement need to be argued and ultimately resolved in any particular circumstance by the parties involved. A common reason is that one or more laboratories have failed to include a particular contribution in their uncertainty budgets or have underestimated the effect of such a contribution.

In any case, the set of measurement standards may well include one or more stated uncertainties which are not credible. It is important that methods of analysis account for this possibility without *a priori* knowledge of which of the set of standards are not credible. There are several such aspects that require attention. Ultimately, the major objective (BIPM, 1999) is

as stated to quantify the degree of equivalence of participating laboratories. Although this report concentrates on a central statistic in this regard, the KCRV, there is no attempt to provide definitive solutions because that can only be done after consideration and debate by the parties concerned. Rather, the intention is to convey an appropriate *attitude* to adopt in the analysis of interlaboratory-comparison data, and to consider some *candidate* analysis tools.

In assigning a KCRV given the results of an interlaboratory experiment, one approach (Pauwels *et al*, 1998) starts by assuming that the differences between individual results, both between and within laboratories, are all of a statistical nature regardless of their causes. The value provided by each laboratory (the laboratory "mean") is considered as an unbiased estimate of the quantity of concern, and usually a weighted or an unweighted mean of the laboratory means is assumed to be the best estimate of the quantity.[3] An uncertainty (under a Gaussian assumption) is then associated with the KCRV. If an analysis indicated that there was no reason to doubt that the laboratory values (with their corresponding uncertainties) could all be regarded as belonging to Gaussian distributions, the uncertainty would be computed using traditional statistics. Otherwise, another approach should be adopted.

The median has been proposed (Müller, 1995) as a more robust estimator of a reference value, because it is less influenced by the presence of extreme values. Further, the MAD estimator (Müller, 1995), also suggested for use in international comparisons, permits the corresponding uncertainty to be estimated.[4]

Let there be N participating laboratories, including the pilot laboratory. The case where a single (scalar) quantity[5] is measured by each laboratory is considered. Specifically, each laboratory provides its best estimate of the measurement standard and of the associated standard uncertainty or expanded uncertainty (ISO, 1995). For j = 1, …, N, let x_j denote the measurement standard provided by Laboratory j and u_j the corresponding standard uncertainty provided by that laboratory.

Several approaches are considered. Within each approach, the starting point is the quoting of a number of assumptions or assertions. The extent to which these assertions apply in any particular instance would need to be examined carefully. Some approaches are based on *ranking* the values of x_j, i.e., by arranging them in ascending order (or in non-descending order if there are ties). It is taken henceforth that the x_j have been so arranged. For simplicity of presentation, the x_j - values are taken as distinct. Minor modification will permit ties.

The discussed approaches are illustrated throughout by a set of 10 measurement standards from a CCPR[6] spectral irradiance comparison (Walker *et al*, 1991). Data was extracted from

[3] A weighted mean would be used if the uncertainties provided with the values could be regarded as credible and an unweighted mean otherwise.

[4] The MAD is the *median absolute deviation* from the median. The product of MAD and an appropriate factor provides an estimate of the standard deviation. The multiplication factor is that which would apply if the underlying distribution were Gaussian. If this assumption did not hold (and this aspect is difficult to test for a small sample), the resulting value might be unreliable. It can, however, be expected to be more robust than the standard deviation determined directly from the data.

[5] Other cases, e.g., where a *vector*, *set* or *sequence* of measurement standards is compared are also important; see Section 8.

[6] CCPR is a Consultative Committee of the CIPM concerned with photometry and radiometry.

the results of this comparison which corresponded to a wavelength of 900 nm. Figure 1 shows these standards and their expanded uncertainties (ISO, 1995) at the 95% level of confidence indicated by horizontal bars. (It also shows, as vertical lines, some statistical results for these standards: see Section 3.) The standards have an arbitrary origin.

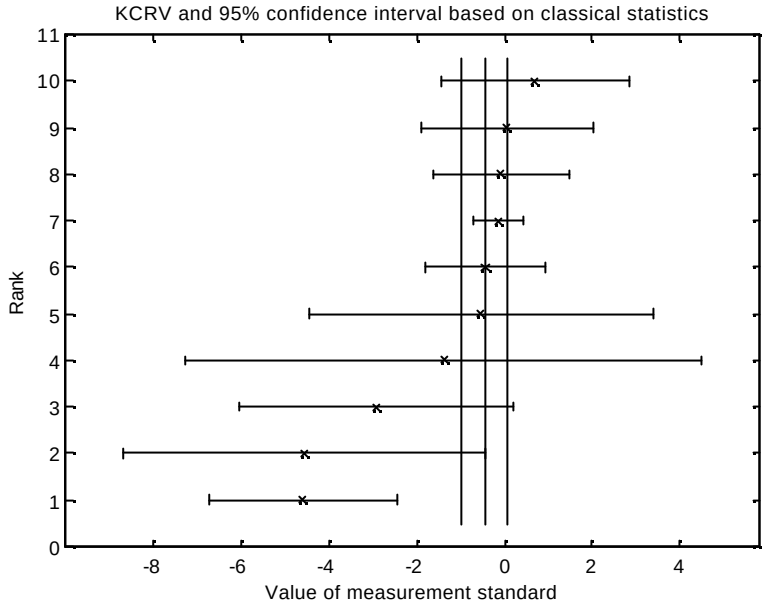

Figure 1. A set of 10 measurement standards from a CCPR spectral-irradiance key comparison. The expanded uncertainties of the standards at the 95% level of confidence are shown (horizontal bars). The standards have been ranked in ascending order. The key-comparison reference value is estimated by the classical weighted mean (central vertical line) and 95% confidence limits (outer vertical lines) are determined under the assumption of normality.

Section 2 lists a set of assertions, some of which would possibly be applicable to a particular key comparison. Section 3 outlines an approach which would be applicable if the quoted uncertainties can all be treated as credible. Section 4 considers the case where the bulk of the quoted uncertainties can be treated as credible. Sections 5 and 6 address the case where it is decided not to accord the quoted uncertainties with any degree of credibility. Section 7 considers the identification of a *cluster* (or subset) of the measurement standards that have a degree of self-consistency as the basis for determining a KCRV. Section 8 contains concluding remarks.

2. Assertions

Some assertions that can be made concerning the provided measurement standards are given. They are not mutually exclusive and are stated informally. The attitude taken here is that, for any one particular comparison, those assertions that are regarded as appropriate are selected, and a suitable statistical approach chosen or derived. It may be necessary to utilize additional or alternative assertions in some circumstances. A major recommendation is that the assertions that lead to a choice of approach are recorded with the results provided by the approach.

The first two assertions below are irrefutable. The remainder have degrees of credibility that must be assessed in any particular situation.

1. A participating laboratory provides a measurement-standard *value* that (generally) is biased. The bias (error) may be positive or negative, i.e., the value may be an overestimate or an underestimate, and may or may not be appreciable (compared with the stated uncertainty, for example).
2. A participating laboratory provides a measurement-standard *uncertainty* that (generally) is biased. The bias may be positive or negative and hence the uncertainty may be pessimistically large or optimistically small.
3. The measurement standards have errors that are mutually independent.
4. For any one laboratory the value bias is equally likely to be positive or negative.
5. The uncertainties can all be regarded as *credible*.
6. *None* of the uncertainties can be regarded as credible.
7. Some of the uncertainties can be regarded as credible, but it is not possible *a priori* to identify which.
8. If, taking account of their values and uncertainties, a sizable subset of the measurement standards forms a *cluster*, those standards in the cluster are regarded as representative, and those outside as unrepresentative.
9. A nominated subset of the values have biases which are expected to have the *same sign* (perhaps, e.g., because a common reference standard is used).
10. The biases of a nominated subset of the measurement standards are expected to have *known signs* (because the measurement method is known to deliver "high" or "low" values), but the magnitudes of the biases are unknown (otherwise the measurement standards could be corrected to account for them).
11. The measurement standards have errors, some of which are mutually independent and the remainder are not, and it is possible to identify which is which.
12. Some of the uncertainties can be regarded as credible, and it is possible to identify which.
13. If some of the measurement standards have errors which are correlated, it is possible to quantify the correlations.

Henceforth, Assertion 3 shall always be deemed to apply, except in Section 8 where other possibilities are briefly reviewed.

3. Credible uncertainties I

Suppose that the standard uncertainties u_j, $j = 1, \ldots, N$, can all be regarded as credible (Assertion 5). The classical solution is that of maximum likelihood and yields the KCRV x_{KCRV} as the weighted mean (Dietrich, 1991, pp39-42) given by

$$\frac{x_{\mathrm{KCRV}}}{u_{\mathrm{KCRV}}^{\,2}} = \sum_{j=1}^{N} \frac{x_j}{u_j^{\,2}},$$

where the standard deviation (standard uncertainty) u_{KCRV} of x_{KCRV} is given by

$$\frac{1}{u_{\mathrm{KCRV}}^{\,2}} = \sum_{j=1}^{N} \frac{1}{u_j^{\,2}}.$$

A 95% confidence interval is given by

$$x_{\mathrm{KCRV}} \pm k u_{\mathrm{KCRV}},$$

where the coverage factor k is taken as the value $t_{N-1,0.95}$ from the *t*-distribution. The application of this classical estimator is illustrated in Figure 1 (in which x_{KCRV} is indicated by the central vertical line and the endpoints of the 95% confidence limits are shown by the outer vertical lines).

4. Credible uncertainties II

In Section 3 the uncertainties of the measurement standards were regarded as credible (Assertion 5). For the classical maximum-likelihood estimator used there, individual uncertainties can have a strong influence on the KCRV. Take a case in which u_1 is very much smaller than the remaining u_j. Then, from the formulae in Section 3, $x_{\mathrm{KCRV}} \cong x_1$ and $u_{\mathrm{KCRV}} \cong u_1$. Thus, a single laboratory which quoted a very small uncertainty would have a dominant effect on the resulting KCRV. If u_1 is realistic the KCRV and its uncertainty so obtained would be correctly influenced by this value. If, on the other hand, u_1 were less reliable than expected, being optimistically small, the effect would be to bias the value of KCRV and make its uncertainty unreasonably small. This situation is recognized, and is a major reason for the move to more robust approaches. In this section attention is paid to the case where the majority of the quoted uncertainties can be regarded as reliable, but *some* are (possibly) unreliable (Assertion 7), but not (obviously) identifiable as such.

One such approach in the presence of unreliable data is the use of the median which, as will be shown in Section 5, arises naturally from making minimal assumptions about the sources of the data. In the approach, as in ISO (1995), each measurement standard is regarded as a *measurement result*, in the form of a Gaussian variable[7] defined by the given value as mean and standard uncertainty as standard deviation. In this context, the set of measurement standards can be regarded as the means of these Gaussian variables or distributions.

An equally valid set of *simulated* measurement standards can be realized by *re-sampling* from these distributions. For the first measurement standard a value is generated at random from the Gaussian distribution with mean x_1 and standard deviation u_1 to produce a new value $x_1^{(1)}$. Similarly, values $x_2^{(1)},\ldots,x_N^{(1)}$ are generated from the corresponding Gaussian distributions for measurement standards 2 to *N*. The median of $x_1^{(1)},\ldots,x_N^{(1)}$ is determined to produce a new estimate $x_{\mathrm{KCRV}}^{(1)}$ of the KCRV. A second sample $x_1^{(2)},\ldots,x_N^{(2)}$ is generated similarly and a further estimate $x_{\mathrm{KCRV}}^{(2)}$ of the KCRV formed by taking *its* median. This process is repeated a large number of times, *M*, say, in all, where *M* is say 50000. The set of median values $x_{\mathrm{KCRV}}^{(1)},\ldots,x_{\mathrm{KCRV}}^{(M)}$ define an *empirical sampling distribution* for x_{KCRV}. The 0.025 and 0.975 quantiles of this set then provide a 95% confidence interval for x_{KCRV}. These quantiles are simply determined as the values in the ordered set $x_{\mathrm{KCRV}}^{(1)},\ldots,x_{\mathrm{KCRV}}^{(M)}$ that occur in positions $0.025M$ and $0.975M$ (interpolating if necessary should these values be non-integral). Figure 2 shows the results obtained when using this approach based on taking $M = 50000$. Section 8 contains a discussion on these results relative to those in other sections.

[7] If the measurement standard is quoted with a finite number of degrees of freedom, the variable should instead be taken as a *t*-distribution with this number of degrees of freedom.

Two points should be made concerning this approach. The first point relates to the extent to which the provided uncertainties could influence the results. They will indeed have an influence, but it can be expected to be less than that for the maximum-likelihood approach of Section 3. This point can be demonstrated using the extreme example above, where u_1 is very small compared with the remaining u_j. In this case all sampled values of x_1 will for practical purposes be essentially equal to the provided value, whereas the other sampled x_j will of course vary according to their assumed distributions. As a consequence the effect of the very small value of u_1 will be to introduce negligible statistical variation in x_{KCRV} from this source. It will, however, introduce a *bias*, which in the worst case could be comparable to the presence of an outlier. However, the median is well-known to have good resilience in the presence of an outlier (Müller, 1995).[8]

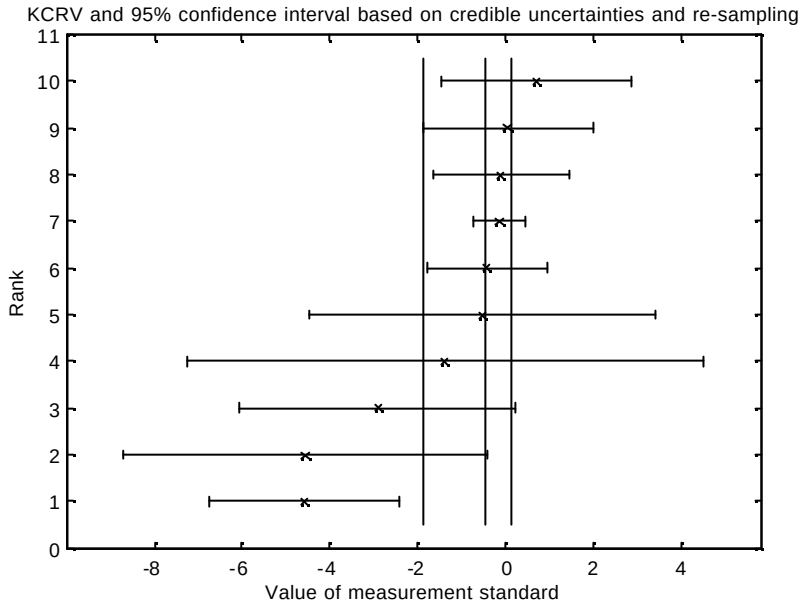

Figure 2. The 10 spectral-irradiance standards and their quoted 95% uncertainties, together with the KCRV (central vertical line) and the endpoints of its 95% confidence interval (outer vertical lines) obtained by re-sampling from the Gaussian distributions implied by the measurement standards.

The second point concerns the choice of the median. It could be argued that the mean could instead be taken, and its empirical sampling distribution similarly determined. However, this distribution tends to normality (Efron, 1982) and hence the results would asymptotically be identical to those for the classical estimator of Section 3. Also see Vecchia and Splett (1994).

5. Non-credible uncertainties I

Suppose that none of the uncertainties u_j can be regarded as credible (Assertion 6) or for political or other reasons it is not possible to nominate which uncertainties can be so regarded. Suppose, however, that for each laboratory the measurement standard has a bias that is equally

[8] The median of a sample having an outlier differs from the median of that sample with the outlier removed by an amount equal to half the difference between the central two ordered standards if N is even, or half the difference between the central standard and one of its immediate neighbours otherwise.

likely to be positive or negative (Assertion 4). An approach based on these assertions is derived and applied.

An estimate $\tilde{x}$ of the KCRV can lie in any one of the following intervals:[9]

Interval No.	0	1	…	$N-1$	N
Interval	$\tilde{x} \le x_1$	$x_1 \le \tilde{x} \le x_2$	…	$x_{N-1} \le \tilde{x} \le x_N$	$x_N \le \tilde{x}$

There is a certain probability that $\hat{x}$ lies in each of these intervals; it is shown how each of these probabilities can be obtained. As a consequence, the most likely interval for $\tilde{x}$ can be determined and, in addition, the intervals containing the endpoints of the 95% confidence interval for $\tilde{x}$ can be found.

Since the errors in the x_j are independent and are equally likely to be positive or negative, the numerical *signs* of these biases follow a *binomial distribution* with probability p = ½ (as do the "number of heads" in a coin-tossing experiment).[10] An estimate $\hat{x}$ is sought which inherits these properties.

Associated with each of the above intervals therefore is a binomial probability. In particular, the probability that $\hat{x}$ lies in Interval k is that of all values of x_j for $j \le k$ being smaller than $\hat{x}$ (and the remaining values greater than $\hat{x}$), which is equal to the binomial probability ${}^{N}C_k / 2^N$.[11]

The probability that $\tilde{x} \le x_k$ is the sum of appropriate binomial probabilities, viz., the probability that $\hat{x}$ lies in the (composite) interval formed by Intervals 0, 1, …, $k-1$, i.e., $\sum_{j=0}^{k-1} {}^{N}C_j / 2^N$. So, for N = 10,

$$\text{Prob}(\tilde{x} \le x_2) = {}^{10}\text{C}_0 / 2^{10} + {}^{10}\text{C}_1 / 2^{10} = \frac{1}{1024} + \frac{10}{1024} = \frac{7}{128} \cong 0.0107.$$

Similarly, again for N = 10,

$$\text{Prob}(\tilde{x} \le x_3) = {}^{10}\text{C}_0 / 2^{10} + {}^{10}\text{C}_1 / 2^{10} + {}^{10}\text{C}_2 / 2^{10} \cong 0.0547.$$

Consider the most likely interval for $\tilde{x}$. From the use of binomial probabilities, this interval is that associated with the "centre" of the (symmetric, bell-shaped) binomial distribution. If N is even, this interval is Interval $N/2$. If N is odd, Interval $(N - 1)/2$ and Interval $(N + 1)/2$ are equally-likely candidates. In the former case, $\hat{x}$ can be chosen as the midpoint of Interval $N/2$. In the latter case, $\hat{x}$ can be chosen as the value $x_{(N+1)/2}$, since it is an endpoint of Interval $(N - 1)/2$ *and* Interval $(N + 1)/2$.

[9] The KCRV could also lie within *two* intervals, viz., at their common boundary. This special case does not affect the main argument here.

[10] This statement is much weaker than one about the *values* of the deviations.

[11] ${}^{N}C_k = N!/\{k!(N-k)!\}$ is the number of ways of choosing k items from N.

The value so determined is the well-known *median* of the N values x_j, $j = 1, \ldots, N$, which is recognized as a robust measure of location in that it has excellent immunity to "wild points" or outliers. It is important to recognize that the median has been *derived* in context here solely as a consequence of the assertions. Müller (1995) advocates the use of the median for interlaboratory comparisons as an approach which is free of subjective decisions. Davies (1988) also favours the median for this purpose, especially in the presence of gross outliers in the data.

A confidence interval $[x_L, x_H]$ associated with the estimator can be obtained as follows. The intervals which contain the endpoints of the 95% confidence interval for $\hat{x}$ are sought. The "natural" choices are for x_L to lie in an interval having endpoints corresponding to cumulative probabilities which bracket the value 0.025, and for x_H to lie in an interval having endpoints corresponding to cumulative probabilities which bracket the value 0.975.[12] A simple and "safe" rule for locating x_L and x_H is inverse linear interpolation in the graph of the corresponding binomial probabilities against the x_j.[13] For $N = 10$, x_L lies between x_2 and x_3, and x_H between x_8 and x_9. Different results will apply for other values of N. Note that there must be at least six standards for this approach to apply: if $N < 6$, $x_L < x_1$ and $x_H > x_N$ and are therefore indeterminate. For N = 6, 7 and 8, $x_1 \le x_L \le x_2$ and $x_{N-1} \le x_H \le x_N$, and can therefore be sensitive functions of the extreme values in the set. It is suggested that N should be nine or more for the application of this approach. The extreme values x_1 and x_N do not influence the determination of the confidence interval for $N \ge 9$.

The approach can be summarized by the following steps:

1. Form the binomial probabilities $p_{N,j} = {}^{N}C_2 / 2^N$, $j = 0, \ldots, N$.
2. Form the cumulative binomial probabilities $P_{N,0} = 0$ and $P_{N,j} = P_{N,j-1} + p_{N,j-1}$, $j = 1, \ldots, N+1$.
3. Associate $\mathrm{P} = \left\{P_{N.j}\right\}_{j=0}^{N+1}$ with the values $\mathrm{X} = \left\{-\infty, x_1, x_2, \ldots, x_N, +\infty\right\}$.
4. Given any P $(0 < P < 1)$, the value of x_p such that the probability that x lies in $(-\infty, x_p)$ is taken as the value given by linear inverse interpolation at P in the table $\left\{\mathrm{X}, \mathrm{P}\right\}$. In particular, $x_{\mathrm{KCRV}} = x_{0.5}$ is the median of X, i.e., of the x_j, and $x_{0.025}$ and $x_{0.975}$ are the endpoints of a 95% confidence interval for x.

Figure 3 shows the cumulative binomial probabilities for N = 10, plotted against the 10 spectral-irradiance standards, and their linear interpolant. The figure also shows the KCRV and its 95% confidence interval, computed as above.

[12] For a symmetric distribution, it would normally be appropriate to define a confidence interval in terms of the 0.025 and 0.975 *quantiles*, as here. In an asymmetric case, the confidence interval could be defined in terms of the q and $1-q$ quantiles, where $0 \le q \le 0.05$. One possibility would be to select q such that the length of this interval was minimized. A benefit would be that the longer tail of the distribution would have less influence on the confidence interval. There may, however, be practical limitations to this choice.

[13] The nature and spacing of the data is such that a more sophisticated interpolation rule, e.g., cubic-spline interpolation, is likely to provide less-reliable results. The reason is that such a rule could lead to spurious oscillations with detrimental effect on the evaluated endpoints.

Note that the distribution as estimated by the 10 standards, using binomial probabilities as above, is decidedly skew. Thus, the estimated confidence interval for the estimator will be asymmetric, and the use of the Gaussian assumption or any assumption of distributional symmetry (such as Gaussian) would lead to invalid results. Figure 4 shows the 10 standards and their quoted 95% uncertainties, together with x_{KCRV} and x_L and x_H. The confidence interval is indeed highly asymmetric. When compared with the use of classical statistics (see Section 3), it is of interest to observe that the estimated values of x_{KCRV} are similar, as are x_H. However, x_L is very different as a consequence of the much longer left-hand tail of the distribution indicated by the standards.[14] The confidence interval is longer than that for classical statistics since the latter implicitly makes stronger assumptions concerning the statistical properties of the data.

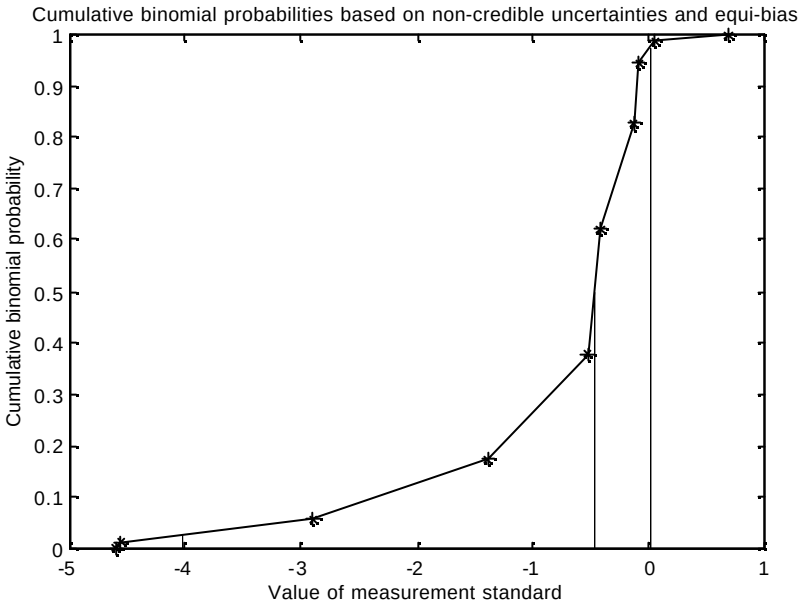

Figure 3. The cumulative binomial probabilities for $N = 10$, plotted against the 10 spectral-irradiance standards, and their linear interpolant. The figure also shows the location of the KCRV (central vertical line) and the endpoints of the 95% confidence interval for the KCRV (outer vertical lines).

Note that the distribution as estimated by the 10 standards, using binomial probabilities as above, is decidedly skew. Thus, the estimated confidence interval for the estimator will be asymmetric, and the use of the Gaussian assumption or any assumption of distributional symmetry (such as Gaussian) would lead to invalid results. Figure 5 shows the 10 standards and their quoted 95% uncertainties, together with x_{KCRV} and x_L and x_H. The confidence interval is indeed highly asymmetric. When compared with the use of classical statistics (see Section 3), it is of interest to observe that the estimated values of x_{KCRV} are similar, as are x_H. However, x_L is very different as a consequence of the much longer left-hand tail of the distribution indicated by the standards.[15] The confidence interval is longer than that for classical statistics since the latter implicitly makes stronger assumptions concerning the statistical properties of the data.

[14] The results for other key-comparison data sets may be very different.

[15] The results for other key-comparison data sets may be very different.

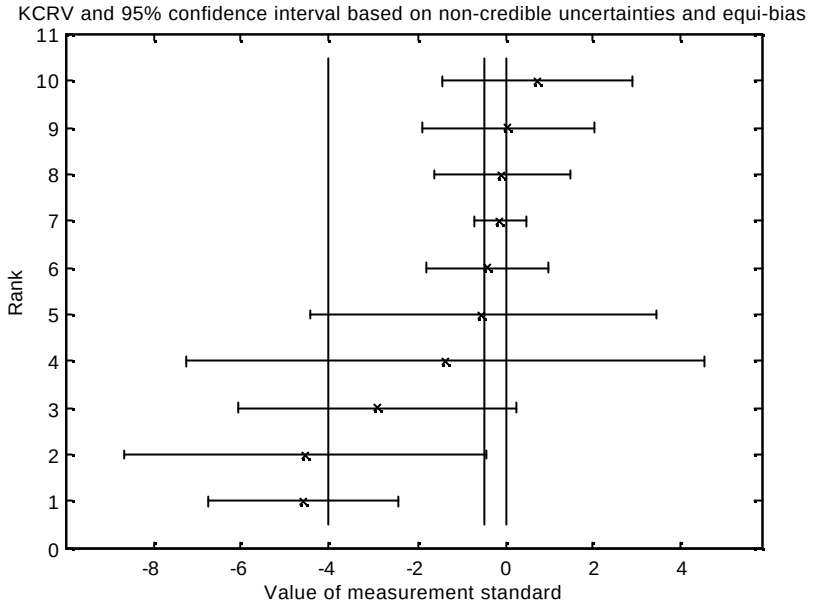

Figure 4. The 10 spectral-irradiance standards and their quoted 95% uncertainties, together with the KCRV (central vertical line) and the endpoints of its 95% confidence interval (outer vertical lines) based on the use of cumulative binomial probabilities.

6. Non-credible uncertainties II

Again suppose, as in Section 5, that none of the uncertainties u_j can be regarded as credible and that the measurement-standard values alone are to be used as the basis of determining a KCRV. In accordance with the comments in Section 1, the set of N measurement standards is taken as *the* population, or at least the knowledge available of the population.

Since the u_j are being disregarded, each x_j is taken as an equally-viable estimate of the value sought. In this sense, it is possible to obtain further valid samples from the existing "sample" of N in order to obtain an indication of the variability of values computed from the sample.

Imagine an infinite number of objects labelled x_1 and an infinite number labelled x_2, and so on. These objects are placed in a box and a random sample of N drawn. The probability of drawing an object labelled x_j, for a prescribed value of j, is evidently $1/N$, in accord with the above statement. By producing many such samples[16] and for each calculating the value of the statistic of interest, e.g., the median, an empirical sampling distribution is determined, as in the approach of Section 4, from which a confidence interval is formed as before. This technique, the bootstrap (Efron, 1982), has in recent times been increasingly used in the physical sciences to

[16] For instance, if $N = 5$ and the standards are therefore x_1, x_2, x_3, x_4 and x_5, the new sample might be x_2, x_1, x_4, x_2 and x_1, or x_1, x_4, x_3, x_5 and x_3.

analyse data. The bootstrap[17] is now being used in metrology (see, e.g., Ciarlini *et al*, 1994; Ciarlini, 1997), and it would seem appropriate to consider its application to interlaboratory-comparison data.

Figure 5 shows the results obtained for the spectral-irradiance standards using 50000 bootstrap samples.

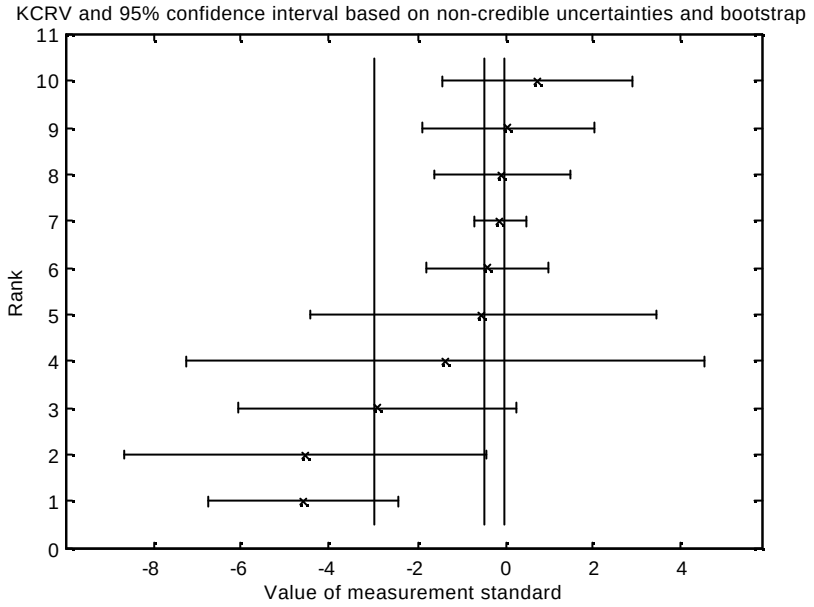

Figure 5. The 10 spectral-irradiance standards and their quoted 95% uncertainties, together with the KCRV (central vertical line) and the endpoints of its 95% confidence interval (outer vertical lines) obtained by direct bootstrap re-sampling from the standards.

7. Clustering

Participating laboratories will provide results of varying quality. Because of the relative degrees of credibility in the quoted uncertainties, a reliable laboratory will provide a 95% confidence interval that is more likely than that from an unreliable laboratory to contain the "correct" value. It follows that results of good quality will tend to be "clustered", i.e., their 95% confidence intervals will tend to "overlap". As a consequence, a reference value can be determined by identifying such a cluster, and utilizing only the measurement standards that comprise it. Therefore, Assertions 7 and 8 are deemed to apply, i.e., that some of the uncertainties can be regarded as credible, but that it is not possible *a priori*, to identify which, and that if a sizable cluster of the standards can be found the standards in the cluster can be regarded as representative.

There are many approaches to clustering (see, e.g., Everitt and Dunn, 1992; Jarvis and Patrick, 1973). One simple way to proceed is as follows. Consider the concept of *candidate* reference values. Every point on the scale of measurement of the quoted values is such a candidate. Some points are more likely than others, and it is required to identify the region or

[17] The bootstrap in the form considered here is usually described as *sampling with replacement*, in which each member of a new sample is drawn from the original sample and then replaced.

regions which are "most likely". For any point on the scale consider the set of 95% confidence intervals that contain it. For points far removed from the quoted values no such confidence interval will embody it. For points near the general "body" of results several such confidence intervals can be expected to contain it. Consider therefore plotting a "consensus diagram" defined as the graph of the number of confidence intervals containing x as a function of x, the variable of interest.

Figure 6 gives an example of a consensus diagram derived from the spectral-irradiance comparison data. All the candidate reference values in the range –0.7 to –0.4 fall within the confidence intervals quoted by (the same) nine of the 10 participating laboratories. For candidate reference values outside this range, eight or fewer laboratories yielded "containing" confidence intervals.

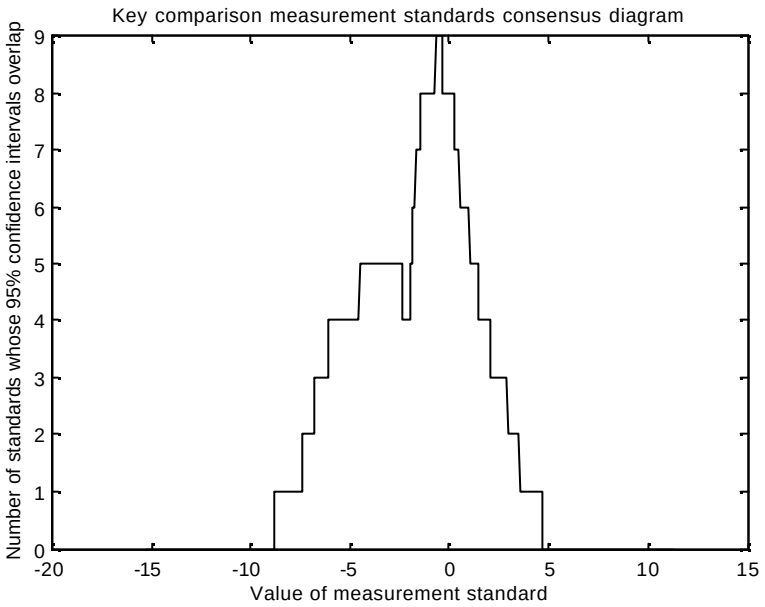

Figure 6. An example of a consensus diagram derived from the spectral-irradiance standards. The graph shows for each candidate position for the reference value the number of participating laboratories whose 95% confidence intervals contain that candidate value.

The next step is to accept the values of the laboratories whose results are so clustered as being more reliable than those outside the cluster.[18] Because the cluster can be regarded as comprising a set of measurement standards that are consistent it is reasonable to determine the maximum-likelihood estimator[19] for the group. This estimator would be given by the formulae in Section 3, but applied only to the standards within the cluster.

[18] There is a risk that some of the less reliable results are contained in the group by chance. Without further information there is little that can be done to ameliorate this risk.

[19] Other estimators such as the median could be used instead.

Figure 7 shows the 10 standards, together with the KCRV computed in this manner from the nine of the 10 standards as above.[20]

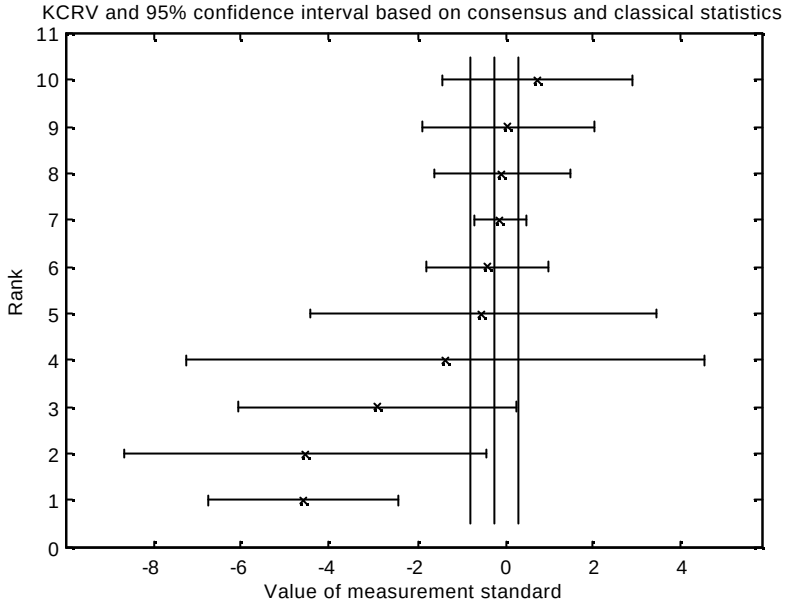

Figure 7. The 10 spectral-irradiance standards, together with KCRV computed from the selected cluster of (nine) standards. The figure also show the endpoints of the 95% confidence interval for the KCRV.

A technique based on overlapping confidence intervals has been applied for computing the uncertainty in an international comparison of fixed points (Pavese, 1984; Pavese *et al*, 1984). However, it was not used for detecting clusters, but to obtain a mean value and a confidence interval.

8. Concluding remarks

An attempt has been made to indicate an appropriate attitude that it is believed should be taken when determining a key-comparison reference value (KCRV). The assertions or assumptions that are believed to apply should be stated explicitly and an analysis technique applied which is consistent with them. Thus, it can be expected that the result obtained will be objective and defensible.

An approach based on such a perception has been illustrated with data from a spectral-irradiance key comparison by making various (sets of) candidate assertions and applying a relevant statistical method. The results obtained, in the form of a KCRV obtained from each technique and the endpoints of a 95% confidence interval for it are summarized numerically in **Error! Reference source not found.**. It is seen that for this data (but not necessarily in general)

1. The stability of the KCRV across approaches is reasonable. The biggest departure is that for the clustering technique (Section 7).

[20] The standard that was not identified as belonging to the cluster of nine (*cf.* Figure 6) is the first in the ranked set.

2. There is a broad range of estimates for the 95% confidence intervals. The upper endpoint of the confidence interval is more stable than the lower, as a consequence of the greater grouping of measurement standards at higher values.
3. The length of the confidence interval varies considerably, the shortest being 1.1 and the longest 4.0. There is a general tendency, as would be expected, for this length to reduce as the assertions made become stronger.

Approach	KCRV	Endpoints of 95% confidence interval	
Credible uncertainties I	-0.4	-1.0	0.1
Credible uncertainties II	-0.5	-1.9	0.2
Non-credible uncertainties I	-0.5	-4.0	0.0
Non-credible uncertainties II	-0.5	-3.0	0.0
Clustering	-0.2	-0.8	0.3

Table 1. Summary statistics for the approaches considered.

It is evident that the results depend on the specific assertions made and utilized. It is regarded as essential that these assertions and the analysis technique employed are recorded with the statement of the KCRV.

The concentration has been on key comparisons relating to a single (scalar) quantity. However, there are key comparisons where a *set* or *sequence* of measurement standards is compared (see below). Also, measurement standards may have correlated errors, perhaps owing to the use of a common reference (also see below). The nature of the assertions made in these cases will be different from those used here.

There are political sensitivities associated with certain statistical techniques. For instance, an alternative approach to those discussed here would attempt to identify outliers (see, e.g., Deutler, 1991), exclude (perhaps some of) them from further consideration, and base a KCRV on the remainder. In the spirit of openness it would be necessary to disclose the measurement standards that had been explicitly excluded, with possibly damaging consequences to all parties concerned. The emphasis here has been different in that all standards in general contribute. It is of relevance to note that the approaches which utilize the median indeed take all standards into account, but extreme values have a relatively weak influence. As implied in Section 4, if a standard having a very large positive bias instead had a very large negative bias, the change in the KCRV would be small, comparable to the spacing of the most "central" standards, and the effect on the confidence interval would be small or zero. It can be argued that the clustering technique (Section 7) is selective. However, since it is more concerned with identifying *consistent* standards it is more benign than outlier-detection techniques but complementary to them.[21]

There are many related aspects that merit attention, including the following.

1. The degree of equivalence of each national measurement standard (Section 1).

[21] Although clustering techniques are not regarded by some statisticians as being as sound as some other approaches, discussions with metrologists on the approach outlined in Section 7 have indicated that it has certain attractive properties for interlaboratory comparisons. The issue requires further consideration and debate.

2. The degree of equivalence between pairs of national measurement standards (Section 1). See, e.g., Wood and Douglas (1996).
3. *Vector-valued* measurement standards, e.g., microwave reflection coefficients having real and imaginary parts in rectangular waveguides (Ridler and Medley, 1996).
4. *Sets* of measurement standards, e.g., a set of gauge blocks (Vaucher, 1995).
5. *Sequences* of measurement standards, e.g., spectrally-varying quantities in which there can be many points, at different wavelengths, for each quantity (Walker *et al*, 1991).
6. Measurement *scales* (Pavese, 2000).
7. The particular problems associated with comparisons involving a small number of laboratories.
8. Comparisons of standards having correlated errors (as, e.g., in radiation-dosimetry key comparisons).
9. Stability of travelling standards used in key comparisons (Frenkel, 1994).
10. The relationship of supplementary comparisons to key comparisons.
11. The statistical efficiency[22] of estimators: those in Section 5 are unbiased but inefficient. (This point was borne out by simulating data from various distributions.)
12. Other statistical estimators such as Huber's skipped mean and Tukey's biweight and outlier tests such as Dixon's and Grubbs', and the use of hypothesis testing. (Davies, 1988).
13. The greater use of statistical modelling (see, e.g., Crowder, 1992; Lischer, 1996). Such approaches, used in analytical and reference-material work, establish a mathematical relationship between the measurements, the reference value, laboratory biases and errors. The model is "fitted" to the measurements and the resulting parameters and their uncertainties used for the purposes considered here.

For some of these aspects the concept of the KCRV is central; for others the KCRV is inessential. Work in the above areas is under way at NPL and elsewhere.

I acknowledge valuable discussions with Patrizia Ciarlini, Peter Harris, Martin Milton, Eulogio Pardo, Franco Pavese, Dave Rayner and Roger Sym.

References

1. BIPM, Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes, BIPM publication (1999), Bureau International des Poids et Mesures, Sèvres, France.
2. Ciarlini, Patrizia *et al*, Non-parametric bootstrap with application to metrological data. In Ciarlini, P. *et al*, *Advanced Mathematical Tools in Metrology* (1994), 219-229, World Scientific, Singapore.
3. Ciarlini, Patrizia, Bootstrap algorithms and applications. In Ciarlini, P. *et al*, *Advanced Mathematical Tools in Metrology* III (1997), 12-23, World Scientific, Singapore.
4. Crowder, M., Interlaboratory comparisons: round robins with random effects. Appl. Statist., **41** (1992), 409-425.
5. Davies, P. L., Statistical evaluation of interlaboratory tests, Fresenius Z. Anal. Chem. **331** (1988), 513-519.

[22] An estimator is efficient if compared with other estimators it has a small mean-squared error.

6. Deutler, T., Grubbs-type estimators for reproducibility variances in an interlaboratory test study, J. Qual, Technol., **23**, (1991), 324-335.

7. Dietrich , C. F., *Uncertainty, Calibration and Probability*, (1991), Adam Hilger, Bristol, UK.

8. Efron, B., *The Jackknife, the Bootstrap, and Other Resampling Plans* (1982), SIAM, Philadelphia.

9. Everitt, B. S. and Dunn, G., *Applied Multivariate Data Analysis* (1992), Oxford University Press, Oxford.

10. Frenkel, R. B., Performance of standards: intercomparison, treatment of correlations, detection of laboratory offsets and outliers, and long-term assessment of an ensemble of standards. In Ciarlini, P. *et al, Advanced Mathematical Tools in Metrology* (1994), 205-217, World Scientific, Singapore.

11. ISO, Guide to the Expression of Uncertainty in Measurement (1995), International Standards Organisation, Geneva, Switzerland.

12. Jarvis, R. A. and Patrick, E.A., Clustering using a similarity measure based on shared nearest neighbours. IEEE Trans. Comput., **C-22** (1973), 1025-1034.

13. Lischer, P., Robust statistical methods in interlaboratory analytical studies. In Riedler, H. (ed.), *Robust Statistics, Data Analysis and Computer Intensive Methods* (1996), 251-265.

14. Müller, J. W., Possible advantages of a robust evaluation of comparisons. Report BIPM-95-2 (1995), Bureau International des Poids et Mesures, Sèvres, France.

15. Pauwels, J., Lamberty, A. and Schimmel, H., The determination of the uncertainty of reference materials certified by laboratory intercomparison. Accred. Qual. Assur. **3** (1998), 180-184.

16. Pavese, F., Final Report of the International intercomparison of fixed points by means of sealed cells: 13.81 K to 90.686 K, BIPM Monograph 84/4 (1984), Bureau International des Poids et Mesures, Sèvres, France).

17. Pavese, F., Mathematical problems in the definition of standards based on scales: the case of temperature. To appear in Ciarlini, P. *et al, Advanced Mathematical Tools in Metrology* (2000), World Scientific, Singapore.

18. Pavese, F., Ancsin, J., Astrov, D. N., Bonhoure, J., Bonnier, G., Furukawa, G. T., Kemp, R. C., Maas, H., Rusby, R. L., Sakurai, H., Shankang, Ling, An international intercomparison of fixed points by means of sealed cells in the range 13.81 K to 90.686 K, *Metrologia* **20** (1984), 127-144.

19. Ridler, N. M. and Medley, J. C., A comparison of microwave reflection and transmission measurements in rectangular waveguides. In Mathur, B. S, *et al* (eds.), *Advances in Metrology and its Role in Quality Improvement and Global Trade* (1996), 364-368.

20. Thomas, C., List of key comparisons, Report BIPM-99/4 (1999), Bureau International des Poids et Mesures, Sèvres, France.

21. Vaucher, B. G. *et al*, European comparison of short gauge block measurement by interferometry. Metrologia **32** (1995), 79-86.

22. Vecchia, D. F. and Splett, J. D., Outlier-resistant methods for estimation and model fitting. In Ciarlini, P. *et al, Advanced Mathematical Tools in Metrology* (1994), 143-154, World Scientific, Singapore.

23. Wood, B. M. and Douglas, R. J., Confidence-interval comparison of a measurement pair for quantifying a comparison. Metrologia **35** (1998), 187-196.

24. Walker, H. *et al*, Results of a CCPR intercomparison of spectral irradiance measurements by national laboratories. J. Res. Natl. Inst. Sci Technol. **96** (1991), 647-668.